# Project StORe: expectations, a solution and some predicted impact from opening up the research data portfolio

**Graham Pryor**

A fundamental role of the university library has long been recognised as the provision of safe custody and the assurance of measured access to the wealth of published scholarship. Traditionally, the library has been synonymous with a collection of books, although an effective twenty-first century library service is more likely to be defined by the extent to which it enables access to information in non-print formats, particularly that which is accessible by electronic means. Yet, whilst the utility of the university library has extended conspicuously to the provision – and interpretation – of digital resources, including most recently the installation of repositories for the preservation and dissemination of research papers, its principal focus has remained upon items or objects that one may consider in some way to have been published, whether in printed or electronic form.

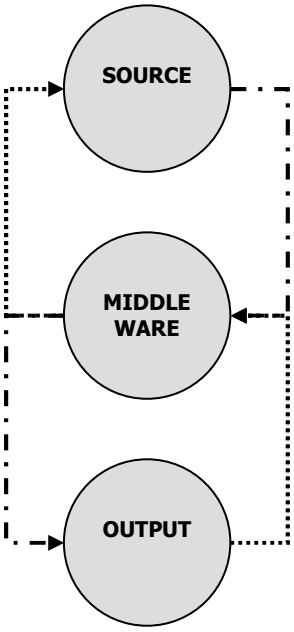
The persistence of such a limited account of library business is perhaps surprising, given the importance vested by universities in their conduct of research and the kudos they perceive it to bestow upon them, since the research output that is visible from published scholarly and scientific articles represents only a fraction of any institution's research undertakings. It is important to ask, therefore, whether libraries might be expected to display a natural interest in the stewardship of all or any of the larger set of 'unpublished' research data that is produced.

While research publications are, for the moment at least, the commodity upon which research performance is assessed, they each serve a resolutely narrow purpose: to present a case and persuade the reader to a particular point of view. As such, they will generally comprise a finely tuned orchestration of theses, arguments and

opinions, developed from a tiny sub-set of data that has been carefully selected and filtered from the much larger accumulation generated. Paradoxically, this larger and predominantly digital collection of research data, from which scholarly articles are eventually derived, constitutes an indisputably valuable asset in its own right, not simply because it is the product of considerable financial and intellectual investment, but for the reason that it has within it the potential for accruing value from further manipulation, analysis and re-use. As with any valuable asset, it deserves a proper mechanism for custodianship and curation that, in tandem with an appropriate level of managed access, will improve on options for maximising the return on investment. When seeking to fulfil this important role, it seems reasonable to postulate that given their several centuries of experience in delivering stewardship for published information, librarians might be called upon to apply their particular portfolio of expertise in sustaining the key intellectual asset that is research data.

If they are waiting for a call to arms they need wait no longer. Currently, the UK's research councils annually invest almost three billion pounds of public money in research, covering the full spectrum of academic disciplines, and increasingly they are concerned that data generated from that research should be managed in ways that better reflect their value. Only this year, the BBSRC and MRC[1] have issued new data policies, both based on principles adopted from the OECD report on *Promoting Access to Public Research Data for Scientific, Economic and Social Development*[2], which recognise that publicly funded research data are a public good, produced in the public interest, and should be openly available to the maximum extent possible. Recognising issues of ownership and intellectual rights, both Councils would allow a limited period of exclusive use, but they require that new research data must be properly curated throughout the information lifecycle and, when released, should include high quality contextual information, or metadata.

Project StORe[3], which has been sponsored by the JISC[4] and CURL[5], was conceived by members of the UK's research library community as an initiative that would apply digital library technologies to create new value for published research. Ostensibly a technical project, its primary objective has been the design of middleware to enable bi-directional links between source repositories containing research data and output repositories containing research publications[6], with the aim of producing direct electronic links between research papers and their source data. Hence, researchers would be able to navigate directly from within an electronic article to the source or synthesised data from which that article was derived. Conversely, direct access would also be provided from source data to the publications associated with those data. The technical challenge of developing this functional middleware was compounded by the diversity of the research data environment, where the StORe project identified sixty-four scientific data types actively being deposited in *source* repositories. Whereas these included images, plots, instrument data, spectra, telemetry, sequences and databases, to name but a small selection, *output* repositories would contain published articles or other texts, usually comprising publications at a pre- or post-refereeing stage, working papers and PhD theses. Output repositories are also frequently referred to as institutional repositories, since they are commonly developed first as an institutional resource, frequently on the basis of a university library initiative. It is also correct to refer to publisher repositories as belonging to the genre. Hence a suite of online periodicals may be considered to be hosted within an output repository.




The relevance of traditional library skills and experience to the design of middleware for bridging across this heterogeneous environment should be apparent. Librarians

comfortable with the digital age have already articulated generic tools for use in other equally diverse contexts: metasearch interfaces to publisher and local databases, metadata harvesters and link resolvers, all based upon recent digital library protocols and standards such as OAI-PMH[7] and qualified Dublin Core[8], with which they have proved themselves adept in providing the kind of high quality contextual information that is now being demanded by the research councils.

To develop a suite of middleware features that would reflect actual needs and aspirations, the project's initial task was to survey researcher behaviours and the processes employed in the generation, organisation and sharing of research data. This survey of seven scientific domains, employing an online questionnaire and one-to-one interviews, was conducted in the Spring of 2006 by a team based at institutions in the UK and USA.

| **Surveying University Library** | **Subject** |
|---|---|
| Edinburgh (lead) / Johns Hopkins | Astronomy |
| Birmingham | Physics |
| Imperial College | Chemistry |
| London School of Economics | Social Sciences |
| Manchester | Biosciences |
| University College London | Biochemistry |
| York (for the White Rose Partnership) | Archaeology |

The team's individual survey reports together provided a comparative topography of the current and potential use of source and output repositories and, after detailed analysis by the project's systems implementers, it proved possible to develop a generic technical specification for the creation of bi-directional links that would directly reflect user requirements. Consequently, a suite of pilot middleware was built and successfully tested in the Social Sciences domain by staff at the UK Data Archive[9] between November 2006 and June 2007. Yet, despite this technical accomplishment, and whilst we were not diverted from our main objective, we found that the survey had opened up a far broader territory than was originally envisaged.

During the survey, researchers had reacted favourably to the opportunities predicted from the putative StORe middleware, with 85% of those who responded declaring that a facility to transfer directly from within an electronic publication to the data upon which its findings are based, or to link instantly to all the publications that have resulted from a particular research dataset, should prove advantageous. This result encouraged us to proceed to the development phase. At the same time, whilst our plans for enhancing the functionality of repositories had originally been laid to improve opportunities for information discovery and data curation, specifically by promising to open a new access route to scientific research data, it became quickly evident that we were also challenging the familiar concept of the academic library, not to mention the very nature of academic publishing.

Traditionally, the publication of research has been understood as the delivery of scholarly output in the form of a printed or electronic document, the process representing a synthesis of large volumes of original and processed data. Most importantly, published papers will have been subject to critical and informed peer review. Subsequently, these papers have been preserved and made available in academic libraries or through electronic portals supported by them. Now, by directly linking these papers to the data from which they were originally derived, the opportunity to explore the basis of a published scholarly argument is at once enlarged and the more detailed background to the testimony of a hypothesis, which previously it was impossible to include in a journal article, is made accessible. Furthermore, the authority of claims made in an article will be more critically assessed when the option to examine the underlying data enables other researchers to compare their own research, data and results.

Of course, making public (or *publishing*) data that has not been through a rigorous process of peer review carries a number of obvious risks. Not least is the potential for invalidating any published paper that follows, if making the data available in

advance is judged to have pre-empted the paper as an original piece of work. This may at first seem like a strong rebuttal of the argument for making data public, but achievable measures for managing such risks do exist, ranging from embargoes and other time constraints on data release to the implementation of robust mechanisms for governing online data upload. More significantly, I would contend that the actual process of making the data public provides its own means of protecting the value of a published paper, since the visibility of its underlying data will serve to improve the quality of the published arguments made within it. In a context where members of the research community at large can access and 'peer review' a paper's source data, few would dare to publish without first being satisfied that awkward questions pertaining to the robustness of conclusions reached might not be raised and pursued!

Nonetheless, the cultural change suggested by this new option to 'publish' research data may be harder to achieve than is implied by such an analysis, and the StORe survey uncovered a realm of attitudes and activities amongst the research community that are not normally exposed to the librarian or information systems provider. Curiously, the strongest messages received were apparently unrelated: a serious necessity to improve upon basic data management practices and the importance of resolving compelling and negative issues of data ownership.

Both the StORe online questionnaire and the series of one-to-one interviews produced evidence of a need for expert assistance with information discovery and organisation, whether this amounted to familiarisation with resources and equipment, in the application of techniques for data organisation and deposit, or with the particular challenge of selecting and assigning metadata. Yet we encountered a general lack of awareness – even resistance – when it came to the availability and use of professional support, which was evident across the seven domains. Notwithstanding the adverse experiences described by researchers, who admitted that metadata assignment was especially demanding in terms of the intellectual effort

required and the burden of time it placed upon them, the development and administration of research data and repositories was not immediately associated with the activities and skills of specialised information intermediaries, albeit their perceived role in data preservation was remarked by respondents to the astronomy survey. Neither could it be established that declared self-reliance led to the practice of good data management. Consistently emphatic in their understanding that the correct assignment of metadata is crucial, and acknowledging a need for assistance from specialists in developing and administering metadata, researchers in all disciplines identified a clear link between the condition of metadata used and the level of support provided by information specialists; but when asked to consider who is responsible for the assignment of metadata to their research output, by far the largest number (212 of the total 377 respondents) claimed that they personally decide which terms to use. Supporting comments indicated that although in a number of cases reference was being made to standard thesauri or schema, this was by no means the norm.

A pervasive culture of self-sufficiency amongst academic researchers goes some way to explaining this general indifference to the role of library or data specialists. Respondents to the StORe survey referred to professional library support having been offered and rejected, expressing a view that it is for researchers themselves to sort out their data problems, and reliance upon documentary or online machine support was consistently preferred to human intervention. As the following table from the questionnaire supports, a majority was found not to seek help when using output repositories (often the province of library professionals) because they perceive there is no assistance available. More disconcerting were the supplementary comments to the table, in which researchers expressed little confidence in what support is provided whilst claiming sufficient comfort with technology to believe themselves equipped to use most IT-enabled services.

| **Question 24.** Do you receive support and/or guidance in your use of output repositories? (This need not take the form of personal support from someone else but could be online prompts, links and advice from within the repositories themselves.) | | | |
|---|---|---|---|
| Documentary support |  | 17.6% | 66 |
| Personal support provided by an intermediary |  | 7.7% | 29 |
| Repository-enabled support |  | 22.1% | 83 |
| No support is provided |  | 28.2% | 106 |
| Unknown |  | 20.7% | 78 |
| Other |  | 3.7% | 14 |

This notorious reluctance to welcome central services to the heart of the research demesne is a barrier that is familiar to many support staff, and the establishment of trust has to be at the core of any attempt to found a productive relationship. Results from the StORe survey also suggest that false assumptions about the notion of *eResearch* and, more specifically for StORe, the underlying principles of *eScience*, may have reinforced this barrier. The number of researchers today who do not use information technology will be insignificant, but whilst many may assume some proficiency it does not make them all accomplished *eResearchers*. The components of this new landscape are twofold: the ubiquitous and ‘always on’ high speed networks, shared infrastructures, cross-community middleware and data standards that together provide a working platform are, for the most part, invisible. Like the power supply to our homes it is there to be switched on, and may be utilised inexpertly and almost at will. But these new technologies and resources will prove advantageous only if they provide researchers with the means of significantly enhancing and improving upon their established research processes and priorities; and in most cases this requires the engagement of information or data professionals with the expertise to bring *eResearch* within the reach of more than a handful of enthusiasts and early adopters.

So who is failing to effect this engagement?

The use of output repositories is one area that has been subject to numerous advocacy campaigns by members of the academic library community, often supported by the availability of training. Yet an attitude commonly encountered during the StORe survey was summed up by one repository user who remarked with some pride that his university had even 'assigned a librarian to our department to help with searches, but I have not used her services'.

The source of this reluctance to engage with professional support is not necessarily trivial and may be found deep in the research tradition. Information technology enables collaboration and the StORe middleware would add to this a new level of openness, but in many areas of the research community the prevailing culture remains one of individual research endeavour. This was underlined when we sought opinion on the methods used for enabling data access and sharing. Most researchers subscribed to the principle of sharing data but there were significant levels of concern about making data available for public access, principally on account of the risks that might contribute to an individual's research profile being usurped.

| **Question 16.** What factors would **discourage** you from sharing your research data? | | |
|---|---|---|
| The threat of loss of ownership | | 202 |
| Risks to an established research niche | | 104 |
| Risk of premature broadcast of research findings | | 235 |
| Subversion of intellectual property rights, including copyright | | 163 |
| Ethical constraints relating to my research | | 58 |
| Consideration of data protection and other confidentiality issues | | 115 |
| The time/effort required to enable sharing | | 193 |
| Risk of diversion from principal objectives through the generation of additional work | | 144 |
| Risk to commercialisation opportunities | | 59 |
| Increased competition for funding | | 77 |

It was no surprise therefore, when we invited comments on the principle of open access that opinions were ambivalent, being determined by the researcher's role as either producer or consumer of the data in question.

In practical terms, methods for data sharing were found to reflect the predominantly private nature of research, with more than 50% relying upon printed or electronic mail to support their efforts toward collaboration, supplemented by the personal exchange of portable media (e.g. CD-roms or USB drives). This seemed to indicate that the application of technology remains subservient to the more traditional and informal understanding of *networking*. The use of open network drives, published URLs and repositories was limited by comparison.

When introducing this discussion of cultural change I referred to two messages received during the StORe survey. Concerns over data ownership have already been discussed, but it might now be appropriate to retract my assertion that certain data management practices found wanting were unrelated. We were of course concerned to discover a very high volume of original and valuable research output being kept on laptops, PC hard drives, CD-roms and other non-networked and inadequately protected storage, and whatever the reason for such practice, some assistance in improving data curation techniques is long overdue. However, when we asked what measures were normally used by researchers to control access to their data, almost one third referred to storage on standalone computers, which may offer some kind of rationale for this unsafe course of action.

Watching developments in 'Big Science' repositories like the Wellcome Trust's *UK PubMed Central* it seems reasonable to suggest that the joint deposit of articles and data is a natural next step in the evolution of publishing. The UK's JISC is already funding several projects in its Digital Repositories programme to explore options for the citation of data, and there are many aspects of research today where knowledge

is represented as data rather than solely in the form of scholarly publications. The human genome project is a prime example. So can this shift in publishing be managed and informed information access be assured, without the dynamics of technological evolution being subdued by the processes that typify publishing?

By enabling bi-directional links between data and publications, Project StORe may be perceived as contributing to the change in scholarly communication practice; but such technological innovations also require the continued presence of knowledge management expertise to ensure that any opportunities they spawn are effectively optimised throughout the information lifecycle. This will not be easy given the barriers already discussed in this paper. StORe's solution, to adopt a Web 2.0 approach already familiar to a cohort of *eResearchers*, enabled us to incorporate both their aspiration to collaborate and their anxiety to protect. It replicates the environment in which the modern researcher interacts both socially and at work, having a structure similar to services like MySpace or Flickr and – most critically – it allows them to remain in control. Using the StORe middleware, researchers decide which of their data items are to be made public or private, they define their collaborations with colleagues or 'friends', and it is for them to choose which items are to be deposited in a repository and made available for publication. Libraries and librarians too must find some means of melding with the research traditions of individual disciplines, in order to provide assistance without usurping the sense of responsibility that researchers have for their research and the data they generate from it. The stewardship of scholarly output may once have seemed naturally to belong with the library, but more recent advances in technology dictate that it now needs to be regained.

---

1 The Biotechnology and Biological Sciences Research Council and the Medical Research Council

2 OECD (2007) *Principles and Guidelines for Access to Research Data from Public Funding* (www.oecd.org/dataoecd/9/61/38500813.pdf)

3 A description of Project StORe can be found at http://jiscstore.jot.com/WikiHome

4 The Joint Information Systems Committee provides support to UK education and research by promoting innovation in new technologies and through the central support of ICT services (www.jisc.ac.uk/)

5 The Consortium of Research Libraries in the British Isles (www.curl.ac.uk/)

6 Source repositories often function as national data centres, such as the *UK Data Archive* (for social sciences) at the University of Essex or the *Archaeology Data Service* at York. Examples of output repositories include the London School of Economics' *Research Articles Online* and the *Edinburgh Research Archive* at the University of Edinburgh

7 The Open Archives Initiative and the OAI Protocol for Metadata Harvesting, which are explained at www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/faq/oai

8 The Dublin Core metadata element set is explained at www.ukoln.ac.uk/metadata/resources/dc

9 The UK Data Archive (www.data-archive.ac.uk) is located at the University of Essex